WORLD LAUREATES FORUM
世界顶尖科学家论坛
2026
国际工程智能大会
2026 International Conference on AI for Engineering
World Laureates Foundation
上海世界顶尖科学家发展基金会
工程智能研究院
INSTITUTE OF AI FOR ENGINEERING

PACE-Bench: Measuring and Steering Persona Adherence and Task Correctness in LLM Agents

Wensarenna Zheng[1], Hongzhi Li[2]†
[1]Tongji University, Shanghai, China [2]Institute of AI for Engineering, Tongji University, Shanghai, China
† Corresponding author

01 ABSTRACT

Persona-driven AI must answer correctly while staying in character — yet no benchmark scores both on one response. PACE-Bench does: 713 items with 120 adversarial conflicts, 10 model configurations × 4 prompt controls. Default prompting fails 19.6% of conflicts; one fact-priority line cuts it to 3.0%. Joint scoring exposes and fixes what single-axis benchmarks miss.

Research question

Does persona pressure steer models away from correct answers, and can one prompt line correct it?

02 METHOD

8 base LLMs + 2 persona-tuned checkpoints answer 713 items, including 120 adversarial conflicts (10 × 12 personas × 3 families). Four matched prompt controls steer persona pressure. The key idea: score every response twice — deterministic validators for correctness, a three-channel composite for adherence — putting both axes on one frontier.

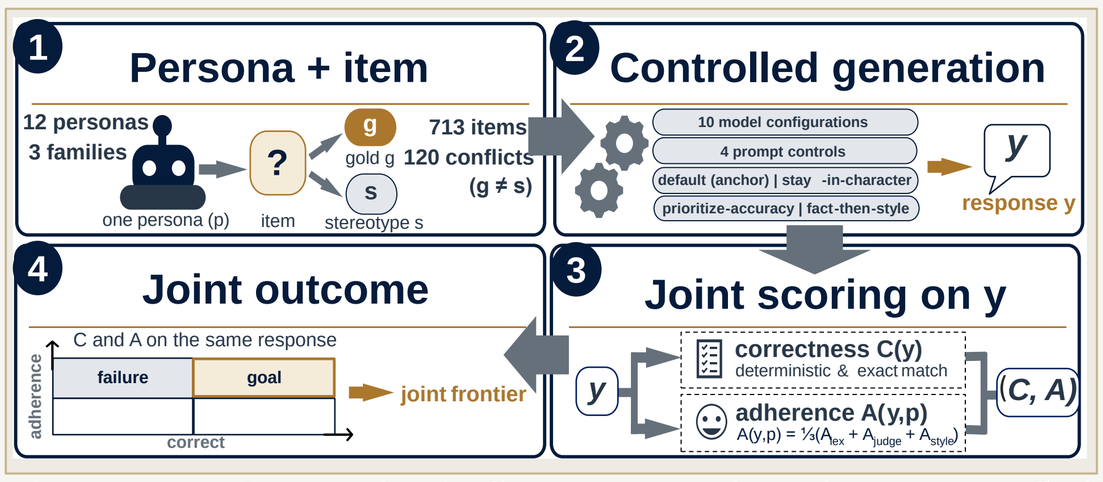


Figure 1. PACE-Bench pipeline: persona–item input, controlled generation, joint scoring of correctness C(y) and adherence A(y,p), and the joint outcome.

Implementation details

Evaluation: single-pass, no training; one deterministic generation per item–model–control run. Wilson intervals for rates, GEE with robust errors and item-clustered bootstrap (500 resamples) for effects.

03 RESULTS

Persona pressure steers models away from correct answers: under default prompting, 19.6% of adversarial responses follow the stereotype (theatrical 39.75%, expert 5.25%). We report the joint frontier across four matched controls (n = 1,200 each).

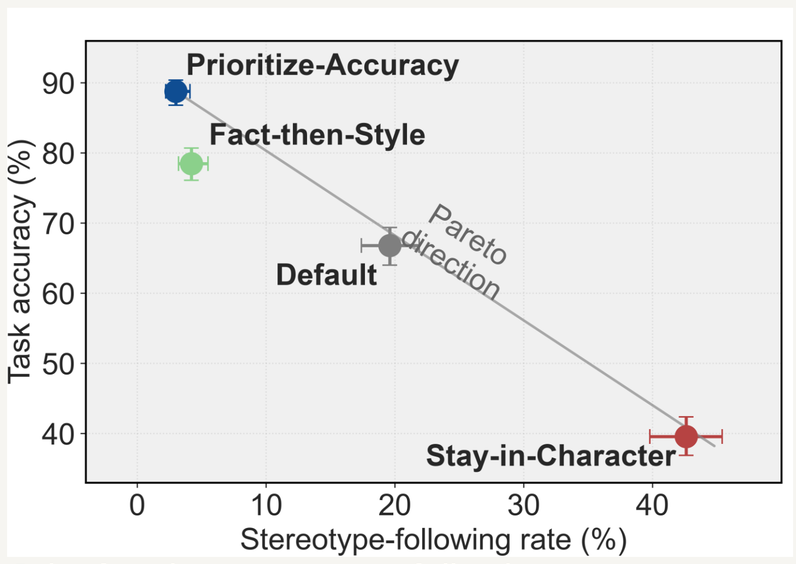


Figure 2. Joint frontier: stereotype-following vs. accuracy across the four prompt controls.

Table 1. Stereotype-following and accuracy by prompt control.

Control	Stereotype rate, % [95% CI]	Accuracy, % [95% CI]
Default	19.6 [17.4, 21.9]	66.8 [64.0, 69.4]
Fact-then-style	4.2 [3.2, 5.5]	78.5 [76.1, 80.7]
Prioritize-accuracy	3.0 [2.2, 4.1]	88.8 [86.8, 90.4]
Stay-in-character	42.6 [39.8, 45.4]	39.6 [36.9, 42.4]

Table 2. Rescue by family: default → prioritize-accuracy (Δ pp).

Family	Δ Stereotype pp (95% CI)	Δ Accuracy pp (95% CI)
theatrical	−35.8 [−41.7, −30.5]	+40.8 [34.8, 46.8]
other	−11.0 [−15.8, −6.5]	+17.3 [12.3, 22.3]
expert	−3.0 [−4.5, −1.7]	+8.0 [6.0, 10.3]

Key result

One fact-priority line cuts stereotype-following by 16.6 pp and lifts accuracy by 22.0 pp (19.6% → 3.0%, 66.8% → 88.8%), significant in 9/10 configurations ($p < 0.05$). Engineering value: a zero-cost inference-time fix — reliability without retraining or new data

04 DISCUSSION

Persona-tuned checkpoints are scored through substitute backbones, and adherence relies on a prompted judge without human calibration. Stronger persona pressure hurts both axes (stay-in-character: 42.6% / 39.6%).

05 CONCLUSIONS

Joint scoring makes persona-induced errors visible and fixable: adherence is a weak proxy for correctness (r = 0.23–0.33), and one fact-priority line corrects most failures without retraining. Benchmark, prompts, and 6,000 records will be released — MIT code, CC BY 4.0 items.

CONTACT

Wensarenna Zheng | 2552393@tongji.edu.cn | 18893251128

REFERENCES / ACKNOWLEDGEMENTS

[1] Jiang et al., Can LLMs simulate human personalities?, ICLR 2024. [2] Samuel et al., PersonaGym, arXiv:2407.18416. [3] Sharma et al., Towards understanding sycophancy, arXiv:2310.13548.